Optimal Index Policies for Quickest Localization of Anomaly in Resource-Constrained Cyber Networks

Kobi Cohen[1], Qing Zhao[1], Ananthram Swami[2]

***Abstract*— We consider the problem of quickest localization of anomaly in a resource-constrained cyber network consisting of multiple components. Due to resource constraints, only one component can be probed at each time. The observations are random realizations drawn from two different distributions depending on whether the component is normal or anomalous. Components are assigned priorities. Components with higher priorities in an abnormal state should be fixed before components with lower priorities to reduce the overall damage to the network. We formulate the problem as a priority-based constrained optimization problem. The objective is to minimize the expected weighted sum of completion times of abnormal components subject to error probability constraints. We then propose a two-stage optimization formulation to solve the problem. First, we consider the independent model, where each component is abnormal independent of other components. Next, we consider the exclusive model, where one only one component is abnormal. We develop optimal index policies under both models. Optimal low-complexity algorithms are derived for the simple hypotheses case, where the distribution is completely known under both hypotheses. Asymptotically (as the error probability approaches zero) optimal low-complexity algorithms are derived for the composite hypotheses case, where there is uncertainty in the distribution parameters. Simulation results then illustrate the performance of the algorithms.**

***Index Terms*—** Anomaly detection, Intrusion Detection System (IDS), sequential hypothesis testing, detection under uncertainty.

I. Introduction

An intrusion detection system (IDS) is a system that monitors the network to detect malicious activities (i.e., attacks) in the network. Once an IDS determines that a malicious activity has occurred, it then alerts the security administrator or initiates a proper response to the malicious activity. Good surveys of IDSs can be found in [1], [2]. Here, we focus on anomaly detection, where statistical methods are used to detect deviations from normal operation. Quickest detection of anomaly subject to reliability constraints is an important requirement when designing intrusion detection schemes. The sooner an IDS detects malicious activities, the lower the resulting damage to the network. Related works of existing techniques for anomaly detection can be found in [3]–[16].

In this paper we address the problem of quickest localization of anomaly in a resource-constrained cyber network. We consider a network with K heterogeneous components which can be paths, routers, or subnets. Assume that an intrusion has been detected. The goal here is to locate the infected components as quickly and as reliably as possible. Most of existing studies on anomaly detection do not consider the constraint on the system monitoring capacity. Here, we focus on a resource-constrained intrusion detection in cyber networks, as was done in [15]–[17]. Due to resource constraints, only one component can be probed at each time. The observations are random realizations drawn from two different distributions depending on whether the component is normal or anomalous. The completion time of component k is defined as the time where the IDS completes testing component k. Components are assigned priorities. Components with higher priorities in an abnormal state should be fixed before those with lower priorities to reduce the overall damage to the network.

Throughout this paper we use the theory of sequential detection. In sequential tests, after each observation has been collected, the detector decides whether to accept H_0, reject H_0 or to take another observation. The sample size achieved by sequential tests can be significantly reduced as compared to fixed-size tests. Therefore, it is a natural approach for quickest localization of anomaly. Sequential detection has been extensively studied in the literature. In cases where the measurements can be collected sequentially according to a specific order, the number of measurements required for optimal detection can be significantly reduced. Related works on this subject can be found in [18]–[21]. However, this is not the case in the IDS model. Change-point detection theory can be applied to the problem of anomaly detection to identify a change in the probability distribution when a malicious activity occurs. Related works on this subject can be found in [8]–[10]. However, in this paper we consider a different problem. Here, an intrusion has been detected (by probing subnet, for instance [15]). The goal here is to locate the infected components. During the anomaly localization, all the observations are drawn from two different distributions depending on whether the component is normal or anomalous. The problem of sequentially testing the simple null hypothesis H_0 versus the simple alternative hypothesis H_1 was solved in [22], [23]. It was shown that the Sequential Probability Ratio Test (SPRT) minimizes the expected sample size under given type I and type II error probability constraints. Related works on SPRT-based solutions for anomaly detection can be found in [3], [5], [6], [13], [14]. Various problems of sequentially testing the composite null hypothesis H_0 versus the composite alternative hypothesis H_1 were studied in [24]–[30]. In this case, asymptotically optimal performance can be obtained as the error probability approaches zero.

[1] Department of Electrical and Computer Engineering, University of California, Davis. Email: {yscohen, qzhao}@ucdavis.edu

[2] Army Research Laboratory, Adelphi, MD 20783. Email: a.swami@ieee.org

This work was supported by Army Research Lab under Grant W911NF1120086

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2013		**00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Optimal Index Policies for Quickest Localization of Anomaly in Resource-Constrained Cyber Networks	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California, Davis,Department of Electrical and Computer Engineering,Davis,CA,95616**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We consider the problem of quickest localization of anomaly in a resource-constrained cyber network consisting of multiple components. Due to resource constraints, only one component can be probed at each time. The observations are random realizations drawn from two different distributions depending on whether the component is normal or anomalous. Components are assigned priorities. Components with higher priorities in an abnormal state should be fixed before components with lower priorities to reduce the overall damage to the network. We formulate the problem as a priority-based constrained optimization problem. The objective is to minimize the expected weighted sum of completion times of abnormal components subject to error probability constraints. We then propose a two-stage optimization formulation to solve the problem. First we consider the independent model, where each component is abnormal independent of other components. Next, we consider the exclusive model, where one only one component is abnormal. We develop optimal index policies under both models. Optimal low-complexity algorithms are derived for the simple hypotheses case, where the distribution is completely known under both hypotheses. Asymptotically (as the error probability approaches zero) optimal low-complexity algorithms are derived for the composite hypotheses case, where there is uncertainty in the distribution parameters. Simulation results then illustrate the performance of the algorithms.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

In the following, we summarize the main results of this paper. We formulate the anomaly localization problem as a priority-based constrained optimization problem. The objective is to minimize the expected weighted sum of completion times of abnormal components (since normal components do not cause damage to the network) subject to error probability constraints. Minimizing the weighted sum of completion times is a natural criterion to prioritize the completion of high-priority components [31]. The optimization is done over the set of all possible selection rules (that the IDS uses to decide which component to test at each time), stopping rules (that the IDS uses to decide when to stop testing each component) and decision rules (that the IDS uses to make a decision regarding the state of each component). We then convert the original optimization problem to a two-stage optimization problem. The the two-stage formulation allows us to simplify the computation of the original optimization problem by decomposing it into two subproblems. We consider both independent and exclusive models. In the former, each component is abnormal, with some prior probability, independent of other components. Under the exclusive model, one and only one component is abnormal with some prior probability (which is a reasonable model when the probability of each component to be compromised is small). We develop index policies under both models. Optimal algorithms are derived for the simple hypotheses case, where the distribution is completely known under both hypotheses. However, in numerous cases under the adversary model, there is uncertainty in the observation distribution (in particular when the component is in an abnormal state). Therefore, we extend our results to the case of composite hypotheses, where there is uncertainty in the distribution parameters. For this case, asymptotically (as the error probability approaches zero) optimal algorithms are derived. In all cases, the algorithms have low-complexity.

The rest of this paper is organized as follows. In Section II we describe the network model and problem formulation. In Section III we present the two-stage optimization formulation. In Sections IV, V we derive optimal low-complexity algorithms under the independent and exclusive models for the simple hypotheses case, respectively. In Section VI we extend our results to derive asymptotically optimal low-complexity algorithms under the independent and exclusive models for the composite hypotheses case. In Section VII we provide applications and numerical examples to illustrate the performance of the algorithms.

II. Network Model and Problem Formulation

Consider a cyber network consisting of K components. Assume that an intrusion has been detected. The goal here is to locate the infected components. Due to resource constraint, only one component can be probed at each time. When component k is tested, a sequence of i.i.d. measurements $\{y_k(i)\}_{i\geq 1}$ is drawn in a one-at-a-time manner. If component k is in a healthy state, $\{y_k(i)\}_{i\geq 1}$ are drawn from distribution $f_k^{(0)}$; if component k is abnormal, $\{y_k(i)\}_{i\geq 1}$ are drawn from distribution $f_k^{(1)}$. We define

$$\mathbf{y}_k(n) = \{y_k(n)\}_{i=1}^{n} \tag{1}$$

as the vector of observations for the n samples that have been collected from component k.

Components are assigned priorities. Let w_k $(0 \leq w_k < \infty)$ be the priority (or weight) of component k. Components with higher priorities in an abnormal state should be fixed before components with lower priorities to reduce the overall damage to the network.

We consider the case where the switching cost is high. Thus, switching between components is done only when testing the current component is completed. The advantages of this scheme are twofold. First, switching between components typically adds significant delay that should be avoided. Second, the IDS is required to store observations of only one component at each time. Thus, this scheme is applicable to limited-memory systems. For convenience, we define t_m as the time where the IDS has completed the $(m-1)^{th}$ test and starts the m^{th} test. After each observation has been collected, the IDS needs to decide whether to take more measurements from the current component or finalize the test on the current component by declaring its state (healthy or abnormal) and choose the next component to test. Let $\pi_k(t_m)$ be the probability (i.e., belief) that component k is abnormal at time t_m. Let $\mathbf{1}_k(t_m)$ be the testing indicator function, where $\mathbf{1}_k(t_m) = 1$ if component k is tested at time t_m and $\mathbf{1}_k(t_m) = 0$ otherwise.

Let N_k be the random sample size required to make a decision regarding the state of component k. Let C_k be the random completion time of testing component k. For example, if the IDS tests component 1 followed by component 2, then $C_1 = N_1$ and $C_2 = N_1 + N_2$.

Let τ_k be a stopping rule, which the IDS uses to decide whether to take more measurements from component k or to finalize the test by declaring its state. Let $\boldsymbol{\tau} = (\tau_1, ..., \tau_k)$ be the vector of stopping rules for the K components.

Let $\delta_k \in \{0, 1\}$ be a decision rule, where $\delta_k = 0$ if the IDS declares that component k is in a healthy state (i.e., H_0), and $\delta_k = 1$ if the IDS declares that component k is in an abnormal state (i.e., H_1). Let $\boldsymbol{\delta} = (\delta_1, ..., \delta_K)$ be the vector of decision rules for the K components.

Let $\phi(t_m) \in \{1, 2, ..., K\}$ be a selection rule, indicates which component is chosen to be tested at time t_m. Let $\boldsymbol{\phi} = (\phi(t_1), ..., \phi(t_K))$ be the vector of selection rules for the K components.

Let

$$\mathcal{H}_1 = \{k : 1 \leq k \leq K \text{ , component } k \text{ is abnormal}\} \text{ ,}$$
$$\mathcal{H}_0 = \{k : 1 \leq k \leq K \text{ , component } k \text{ is healthy}\} \text{ ,}$$

be the sets of all the abnormal and healthy components, respectively.

The problem is to find a selection rule $\boldsymbol{\phi}$, a stopping rule $\boldsymbol{\tau}$ and a decision rule $\boldsymbol{\delta}$ that minimize the expected weighted sum of completion times of all the abnormal components subject

to error probability constraints for each component:

$$\begin{aligned} \inf_{\boldsymbol{\tau},\boldsymbol{\delta},\boldsymbol{\phi}} \quad & \mathbf{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k\right\} \\ s.t. \quad & P_k^{FA} \le \alpha_k \qquad \forall k=1,...,K\ , \\ & P_k^{MD} \le \beta_k \qquad \forall k=1,...,K\ . \end{aligned} \tag{2}$$

Higher penalties are assigned to higher-priority components in an abnormal state [1]. No penalty is associated with components in a healthy state since they do not cause damage to the network. Note that the policy $(\boldsymbol{\phi},\boldsymbol{\tau},\boldsymbol{\delta})$ is dynamic. At each time, the IDS needs to decide whether to take more measurements from component k or to finalize the test by declaring its state and select the next component.

Throughout this paper we develop optimal and asymptotically optimal algorithms to solve (2) under the simple and composite hypotheses cases, respectively. The algorithms developed throughout this paper can be applied to other network models as well. We discuss these extensions in Section VIII.

III. Two-stage Optimization Problem

Instead of solving (2) directly, we propose a two-stage optimization problem. At the first stage, the problem is to find a stopping rule τ_k and a decision rule δ_k for every component k that minimize the expected sample size given H_i subject to error probability constraints:

$$\begin{aligned} \inf_{\tau_k,\delta_k} \ & E(N_k|H_i)\ , \quad i=0,1 \\ s.t. \quad & P_k^{FA} \le \alpha_k\ , \\ & P_k^{MD} \le \beta_k\ . \end{aligned} \tag{3}$$

For the simple hypotheses case, the solution to the first-stage optimization problem (3) is given by the SPRT [22], [23]. Let

$$L_k(n) = \frac{\prod_{i=1}^n f_k^{(1)}(y_k(i))}{\prod_{i=1}^n f_k^{(0)}(y_k(i))} \tag{4}$$

be the Likelihood Ratio (LR) between the two hypotheses of component k at stage n.

Let A_k, B_k $(B_k > 1/A_k)$ be the boundary values used by the SPRT for component k, such that the error constraints are satisfied[2]. According to the SPRT algorithm, at each stage n, the LR is compared to the boundary values as follows:

- If $L_k(n) \in \left((A_k)^{-1}, B_k\right)$, continue to take observations from component k.
- If $L_k(n) \ge B_k$, stop taking observations from component k and declare it as abnormal (i.e., H_1). Clearly, $N_k = n$.
- If $L_k(n) \le (A_k)^{-1}$, stop taking observations from component k and declare it as normal (i.e., H_0). Clearly, $N_k = n$.

Remark 1*:* Implementing sequential tests requires to compute boundary values to determine the stopping rule, such that error constraints are satisfied. In general, the exact determination of the boundary values is very laborious and depends on the observation distribution. However, since the solution to (3) is given by the SPRT, Wald's approximation can be applied to simplify the computation [23]:

$$B_k \approx \frac{1-\beta_k}{\alpha_k}\ , A_k \approx \frac{1-\alpha_k}{\beta_k}\ . \tag{5}$$

Wald's approximation performs well for small α_k, β_k. Since type I and type II errors are typically small, Wald's approximation is widely in practice [23].

For the composite hypotheses case, where there is uncertainty in the distribution parameters, we can obtain asymptotically optimal solution to (3). This case is discussed in Section VI.

At the second stage, the problem is to find a selection rule $\boldsymbol{\phi}$ that minimizes the objective function, given the solution to the K subproblems (3):

$$\begin{aligned} \inf_{\boldsymbol{\phi}} \quad & \mathbf{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k\right\} \\ s.t. \quad & \text{solutions to (3) are given for } k=1,...,K\ . \end{aligned} \tag{6}$$

The solutions to the second-stage optimization problem for the independent and exclusive models are given in Sections IV and V, respectively.

The formulation of the two-stage optimization problem allows us to decompose the original optimization problem (2) into $K+1$ subproblems (3) and (6). We use this formulation to design the solution to (2). In subsequent sections we show that for the simple hypotheses case the solution to the two-stage optimization problem solves the original optimization problem (2) under both independent and exclusive models. For the composite hypotheses case, the solution to the two-stage optimization problem asymptotically (as the error probability approaches zero) solves the original optimization problem under both independent and exclusive models.

IV. The independent Model Case

In this section we consider the independent model under the simple hypotheses case. Under the independent model, each component is abnormal independent of other components. The posterior probability of component k being abnormal can be updated at time t_{m+1} as follows:

$$\begin{aligned} \pi_k(t_{m+1}) = & \\ & \frac{\mathbf{1}_k(t_m)\pi_k(t_m) f_k^{(1)}(\mathbf{y}_k(N_k))}{\pi_k(t_m) f_k^{(1)}(\mathbf{y}_k(N_k)) + (1-\pi_k(t_m))\, f_k^{(0)}(\mathbf{y}_k(N_k))} \\ & \qquad + (1-\mathbf{1}_k(t_m))\,\pi_k(t_m)\ . \end{aligned} \tag{7}$$

In the following we derive optimal low-complexity algorithm for this case.

A. The Proposed Solution

We use the two-stage optimization problem to design the solution to (2). For the simple hypotheses case, the solution to the first-stage optimization problem (3) is given by the SPRT,

[1]Note that the loss due to missed-detection events is negligible for small error probability, since $P_k^{MD} \in O(1/B_k)$ and $E(N_k) \in \Theta(\log B_k)$, where B_k is a boundary value of the sequential test [23], [27].

[2]We discuss the determination of the boundary values A_k, B_k in Remark 1.

discussed in section III. Thus, here we focus on the solution to the second-stage optimization problem (6).

It was shown in [31] that the optimal selection rule for the problem of minimizing the expected weighted sum of completion times given the expected testing time of each component is to select the components in decreasing order of $w_k/E(N_k)$. However, the problem in (6) is different. First, the objective is to minimize the expected weighted sum of completion times of abnormal components only. Second, the expected sample size depends on the component state. In what follows we derive a modified optimal selection rule that solves the second-stage optimization problem (6).

Theorem 1: Let $E(N_k)$ be the solution to (3). A selection-rule $\boldsymbol{\phi}^*$ that selects the components in decreasing order of $\pi_k(t_1)w_k/E(N_k)$ solves the second-stage optimization problem (6).

Proof: The theorem follows from the proof of Theorem 2. ■

Remark 2: The solution to the second-stage optimization problem (6) requires one to compute the expected sample size $E(N_k)$ for all $k = 1, 2, ..., K$ to select the components in decreasing order of $\pi_k(t_1)w_k/E(N_k)$. In general, it is difficult to obtain a closed-form expression to $E(N_k)$. However, since the solution to (3) is given by the SPRT, Wald's approximation can be applied to simplify the computation [23]. For every $i, j = 0, 1$, let

$$D_k(i||j) = E_i\left(\log\frac{f_k^{(i)}(y_k(1))}{f_k^{(j)}(y_k(1))}\right) \tag{8}$$

be the Kullback-Leibler (KL) divergence between the hypotheses H_i and H_j, where the expectation is taken with respect to $f_k^{(i)}$.

The expected sample size is well approximated by [23]:

$$\begin{aligned} E(N_k|H_0) &\approx \frac{(1-\alpha_k)\log\tilde{A}_k - \alpha_k\log\tilde{B}_k}{D_k(0||1)}\,, \\ E(N_k|H_1) &\approx \frac{(1-\beta_k)\log\tilde{B}_k - \beta_k\log\tilde{A}_k}{D_k(1||0)}\,, \end{aligned} \tag{9}$$

where $\tilde{A}_k = (1-\alpha_k)/\beta_k, \tilde{B}_k = (1-\beta_k)/\alpha_k$ are the approximation to A_k, B_k, given in (5).

Thus, at each time t, the expected sample size required to make a decision regarding the state of component k is given by:

$$E(N_k) = \pi_k(t)E(N_k|H_1) + (1-\pi_k(t))E(N_k|H_0)\,, \tag{10}$$

where the approximation approaches the exact expected sample size for small α_k, β_k. Since type I and type II errors are typically small, Wald's approximation is widely used in practice. [23].

Based on the solution to the two-stage optimization problem, we propose Algorithm 1, presented in Table I, to solve (2). Sorting the components in step 1 can be done in $O(k\log k)$ time via sorting algorithms. Then, a series of SPRTs is performed according to this order until all the components are tested. The index policy described in Algorithm 1 is intuitively satisfying. The priority of component k in terms of testing order should be higher as the weight w_k increases, or the probability to be abnormal $\pi_k(t_1)$ increases, or the expected sample size $E(N_k)$ decreases (since $E(N_k)$ is added to the completion time of every component which is tested after component k). The SPRT is used to minimize the expected sample size to reduce the completion times.

TABLE I

ALGORITHM 1 FOR THE INDEPENDENT MODEL

1. arrange the components in decreasing order of $\pi_k(t_1)w_k/E(N_k)$
2. for $k = 1, ..., K$ components do:
3. perform SPRT for component k, with $P_k^{FA} \le \alpha_k$, $P_k^{MD} \le \beta_k$
4. end for

B. Optimality of Algorithm 1

In this section we provide performance analysis of Algorithm 1. Note that Algorithm 1 uses a static selection rule (as stated in step 1), where the components order is predetermined at time t_1. However, the performance analysis in this section is not restricted to static selection rules. The following theorem shows that Algorithm 1 is optimal among the class of both static and dynamic selection rules (that update the selection dynamically at each time t_k).

Theorem 2: Under the independent model, Algorithm 1 solves (2).

Proof: Let $E'(N_k|H_i,\mathrm{t})$ be the expected sample size achieved by a stopping rule and a decision rule $(\tau_k'(t), \delta_k'(t))$, depending on the time that component k is tested (i.e., $(\tau_k'(t), \delta_k'(t))$ depend on the selection rule), such that error constraints are satisfied. Let $E^{A1}(N_k|H_i)$ be the expected sample size achieved by the SPRT's stopping rule and decision rule $(\tau_k^{A1}, \delta_k^{A1})$, independent on the time that component k is tested (i.e., $(\tau_k^{A1}, \delta_k^{A1})$ are independent on the selection rule), such that error constraints are satisfied. Clearly, $E^{A1}(N_k|H_i) \le E'(N_k|H_i, t)$ for all k, t, for $i = 0, 1$ and are achieved by Algorithm 1.

First, consider the case where $K = 2$. Assume that

$$\frac{\pi_1(t_1)w_1}{E^{A1}(N_1)} \ge \frac{\pi_2(t_1)w_2}{E^{A1}(N_2)}\,.$$

Consider selection rules $\boldsymbol{\phi}^{(1)}$, $\boldsymbol{\phi}^{(2)}$ that select component 1 first followed by component 2 and component 2 first followed by component 1, respectively. The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned} &\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(2)})\right\} \\ &\quad = \left(E'(N_2|H_1, t_1)\right)\pi_2(t_1)w_2 \\ &\qquad + \left(E'(N_2|t_1) + E'(N_1|H_1, t_2)\right)\pi_1(t_1)w_1. \end{aligned} \tag{11}$$

The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(1)})\right\}\\ &\quad= \left(E'(N_1|H_1,t_1)\right)\pi_1(t_1)w_1\\ &\qquad+\left(E'(N_1|t_1)+E'(N_2|H_1,t_2)\right)\pi_2(t_1)w_2.\end{aligned} \tag{12}$$

Note that the expected weighted sum of completion times achieved by both selection rules can be further reduced by minimizing the expected sample sizes (such that error constraints are satisfied) independent on the selection rules, which achieved by $(\tau_k^{A1}, \delta_k^{A1})$. Therefore, an optimal solution has to be $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})$ or $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(2)})$. Next, we use the interchange argument to prove the theorem for $K = 2$. The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(2)})\right\}\\ &\quad= \left(E^{A1}(N_2|H_1)\right)\pi_2(t_1)w_2\\ &\qquad+\left(E^{A1}(N_2)+E^{A1}(N_1|H_1)\right)\pi_1(t_1)w_1.\end{aligned} \tag{13}$$

The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})\right\}\\ &\quad= \left(E^{A1}(N_1|H_1)\right)\pi_1(t_1)w_1\\ &\qquad+\left(E^{A1}(N_1)+E^{A1}(N_2|H_1)\right)\pi_2(t_1)w_2.\end{aligned} \tag{14}$$

The expected weighted sum of completion times achieved by $\boldsymbol{\phi}^{(1)}$ is lower than the expected weighted sum of completion times achieved by $\boldsymbol{\phi}^{(2)}$ since that $\frac{\pi_1(t_1)w_1}{E^{A1}(N_1)} \geq \frac{\pi_2(t_1)w_2}{E^{A1}(N_2)}$, which completes the proof for $K = 2$.

Next, we prove the theorem by induction on the number of components K. Assume that the theorem is true for $K-1$ components. Assume that

$$\frac{\pi_1(t_1)w_1}{E^{A1}(N_1)} \geq \frac{\pi_2(t_1)w_2}{E^{A1}(N_2)} \geq ... \geq \frac{\pi_K(t_1)w_K}{E^{A1}(N_K)}.$$

Consider an optimal selection rule $\boldsymbol{\phi}^{(j)}$ that selects component j first. Due to the independency between components, it can be verified by the induction hypothesis that the last $K-1$ components have to be selected in decreasing order of $\pi_k(t_1)w_k/E^{A1}(N_k)$ and tested by the SPRT. Hence, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(j)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(j)})\right\}\\ &= \pi_j(t_1)w_j\left(E'(N_j|H_1,t_1)\right)\\ &+\sum_{k=1,k\neq j}^{K}\Bigg[\pi_k(t_1)w_k\times\\ &\left(E'\left(N_j|t_1\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A1}\left(N_i\right)\right)+E^{A1}\left(N_k|H_1\right)\right)\Bigg].\end{aligned} \tag{15}$$

First, note that the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(j)})$ can be further reduced for all j by minimizing the expected sample size $E'(N_j|H_i,t_1)$ for $i = 0, 1$, which achieved by $(\tau_j^{A1}, \delta_j^{A1})$. Therefore, an optimal solution has to be $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(j)})$ for an optimal selection rule $\boldsymbol{\phi}^{(j)}$. Thus, in the following we consider solutions of the form $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi})$.

Next, by contradiction, consider an optimal selection rule $\boldsymbol{\phi}^{(j\neq1)}$ that selects component $j \neq 1$ first. Therefore, $\boldsymbol{\phi}^{(j\neq1)}$ selects the components by the following order:

$$j, 1, 2, ..., j-1, j+1, ..., K.$$

As a result, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(j\neq1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(j\neq1)})\right\}\\ &= \pi_j(t_1)w_j\left(E^{A1}(N_j|H_1)\right)\\ &\quad+\pi_1(t_1)w_1\left[E^{A1}\left(N_j\right)+E^{A1}\left(N_1|H_1\right)\right]\\ &\quad+\sum_{k=2,k\neq j}^{K}\Bigg[\pi_k(t_1)w_k\times\\ &\left(E^{A1}\left(N_j\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A1}\left(N_i\right)\right)+E^{A1}\left(N_k|H_1\right)\right)\Bigg].\end{aligned} \tag{16}$$

We use the interchange argument to prove the theorem. Consider a selection rule $\boldsymbol{\phi}^{(1)}$ that selects component 1 first followed by components $j, 2, 3, j-1, j+1, ..., K$. Similar to (16), the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})\right\}\\ &= \pi_1(t_1)w_1\left(E^{A1}(N_1|H_1)\right)\\ &\quad+\pi_j(t_1)w_j\left[E^{A1}\left(N_1\right)+E^{A1}\left(N_j|H_1\right)\right]\\ &\quad+\sum_{k=2,k\neq j}^{K}\Bigg[\pi_k(t_1)w_k\times\\ &\left(E^{A1}\left(N_j\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A1}\left(N_i\right)\right)+E^{A1}\left(N_k|H_1\right)\right)\Bigg].\end{aligned} \tag{17}$$

By comparing (16) and (17), it can be verified that:

$$\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(1)})\right\}$$
$$\leq \mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A1}, \boldsymbol{\delta}^{A1}, \boldsymbol{\phi}^{(j\neq 1)})\right\}$$

since that $\pi_1(t_1)w_1/E^{A1}(N_1) \geq \pi_j(t_1)w_j/E^{A1}(N_j)$.
The expected weighted sum of completion times can be reduced by selecting component 1 first followed by component j, which contradicts the optimality of $\boldsymbol{\phi}^{(j\neq 1)}$. Hence, at time t_1 selecting component 1 minimizes the expected weighted sum of completion times. By the induction hypothesis, for the last $K-1$ components we select the components in decreasing order of $\pi_k(t_1)w_k/E^{A1}(N_k)$, which completes the proof. ■

V. The Exclusive Model Case

In this section we consider the exclusive model under the simple hypotheses case. Under the exclusive model, one and only one component is abnormal. The posterior probability of component k being abnormal is updated at time t_{m+1} as given in (18) on the next page. It is easy to see that under the exclusive model, we have $\sum_{k=1}^{K}\pi_k(t) = 1$. Note that in contrast to the independent model, under the exclusive model the beliefs of all the components are changed at each time due to the dependency across components. The posterior probabilities depend on the selection rule and the collected measurements. Nevertheless, in what follows we propose an optimal low-complexity algorithm to solve (2) based on the two-stage optimization problem (3), (6). In section V-B we provide an optimality analysis.

A. The Proposed Solution

We use the two-stage optimization problem to design the solution to (2). For the simple hypotheses case, the solution to the first-stage optimization problem (3) is given by the SPRT, discussed in section III. Thus, here we focus on the solution to the second-stage optimization problem (6). In section IV-A, we showed that selecting the components in decreasing order of $\pi_k(t_1)w_k/E(N_k)$ solves (6) under the independent model. In the following we show that a different selection rule solves (6) under the exclusive model.

Theorem 3: Let $E(N_k|H_i), i = 0, 1$ be the solution to (3). A selection rule $\boldsymbol{\phi}^*$ that selects the components in decreasing order of $\pi_k(t_1)w_k/E(N_k|H_0)$ solves the second-stage optimization problem (6).

Proof: The theorem follows from the proof of Theorem 4. ■

Based on the solution to the two-stage optimization problem, we propose Algorithm 2, presented in Table II, to solve (2). The index policy described in Algorithm 2 is intuitively satisfying. The priority of component k in terms of testing order should be higher as the weight w_k increases, or the probability to be abnormal $\pi_k(t_1)$ increases, or the expected sample size $E(N_k|H_0)$ decreases. Note that in contrast to the independent model, here we take into account the expected sample size under H_0 solely. The reason is that if component k is abnormal, there is no penalty to other components under the exclusive model (since only one component is abnormal).

TABLE II
Algorithm 2 for the exclusive model

1. arrange the components in decreasing order of $\pi_k(t_1)w_k/E(N_k|H_0)$
2. for $k = 1, ..., K$ components do:
3. perform SPRT for component k, with $P_k^{FA} \leq \alpha_k$, $P_k^{MD} \leq \beta_k$
4. end for

On the other hand, if component k is healthy, then $E(N_k|H_0)$ is added to the completion time of the components which are tested after component k (and may be abnormal). The SPRT is used to minimize the expected sample size to reduce the completion times.

B. Optimality of Algorithm 2

In this section we provide performance analysis of Algorithm 2. Note that Algorithm 2 uses a static selection rule (as stated in step 1), where the components order is predetermined at time t_1. However, the performance analysis in this section is not restricted to static selection rules. The following theorem shows that Algorithm 2 is optimal among the class of both static and dynamic selection rules (that update the selection dynamically at each time t_k).

Theorem 4: Under the exclusive model, Algorithm 2 solves (2).

Proof: Let $E'(N_k|H_i,\text{t})$ be the expected sample size achieved by a stopping rule and a decision rule $(\tau_k'(t), \delta_k'(t))$, depending on the time that component k is tested (i.e., $(\tau_k'(t), \delta_k'(t))$ depend on the selection rule), such that error constraints are satisfied. Let $E^{A2}(N_k|H_i)$ be the expected sample size achieved by the SPRT's stopping rule and decision rule $(\tau_k^{A2}, \delta_k^{A2})$, independent on the time that component k is tested (i.e., $(\tau_k^{A2}, \delta_k^{A2})$ are independent on the selection rule), such that error constraints are satisfied. Clearly, $E^{A2}(N_k|H_i) \leq E'(N_k|H_i, t)$ for all k, t, for $i = 0, 1$.

First consider the case where $K = 2$. Assume that

$$\frac{\pi_1(t_1)w_1}{E^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)w_2}{E^{A2}(N_2|H_0)} . \tag{19}$$

Consider selection rules $\boldsymbol{\phi}^{(1)}$, $\boldsymbol{\phi}^{(2)}$ that select component 1 first followed by component 2 and component 2 first followed by component 1, respectively. The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(2)})$ is given by:

$$\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(2)})\right\}$$
$$= \left(E'(N_2|H_1, t_1)\right)\pi_2(t_1)w_2$$
$$+ \left(E'(N_2|H_0, t_1) + E'(N_1|H_1, t_2)\right)\pi_1(t_1)w_1. \tag{20}$$

$$\pi_k(t_{m+1}) = \frac{\mathbf{1}_k(t_m)\pi_k(t_m)f_k^{(1)}(\mathbf{y}_k(N_k))}{\pi_k(t_m)f_k^{(1)}(\mathbf{y}_k(N_k)) + (1-\pi_k(t_m))\, f_k^{(0)}(\mathbf{y}_k(N_k))} + \frac{(1-\mathbf{1}_k(t_m))\,\pi_k(t_m)f_{\phi(t_m)}^{(0)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)}))}{\pi_{\phi(t_m)}(t_m)f_{\phi(t_m)}^{(1)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)})) + \left(1-\pi_{\phi(t_m)}(t_m)\right) f_{\phi(t_m)}^{(0)}(\mathbf{y}_{\phi(t_m)}(N_{\phi(t_m)}))} \,. \tag{18}$$

The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t), \boldsymbol{\delta}'(t), \boldsymbol{\phi}^{(1)})\right\} \\ &= (E'(N_1|H_1,t_1))\,\pi_1(t_1)w_1 \\ &\quad + (E'(N_1|H_0,t_1) + E'(N_2|H_1,t_2))\,\pi_2(t_1)w_2.\end{aligned} \tag{21}$$

Note that the expected weighted sum of completion times achieved by both selection rules can be further reduced by minimizing the expected sample sizes (such that error constraints are satisfied) independent on the selection rules, which achieved by $(\tau_k^{A2}, \delta_k^{A2})$. Therefore, an optimal solution has to be $(\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(1)})$ or $(\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(2)})$. Next, we use the interchange argument to prove the theorem for $K = 2$. The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(2)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(2)})\right\} \\ &= \left(E^{A2}(N_2|H_1)\right)\pi_2(t_1)w_2 \\ &\quad + \left(E^{A2}(N_2|H_0) + E^{A2}(N_1|H_1)\right)\pi_1(t_1)w_1.\end{aligned} \tag{22}$$

The expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2}, \boldsymbol{\delta}^{A2}, \boldsymbol{\phi}^{(1)})\right\} \\ &= \left(E^{A2}(N_1|H_1)\right)\pi_1(t_1)w_1 \\ &\quad + \left(E^{A2}(N_1|H_0) + E^{A2}(N_2|H_1)\right)\pi_2(t_1)w_2.\end{aligned} \tag{23}$$

The expected weighted sum of completion times achieved by $\boldsymbol{\phi}^{(1)}$ is lower than the expected weighted sum of completion times achieved by $\boldsymbol{\phi}^{(2)}$ since that $\frac{\pi_1(t_1)w_1}{E^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)w_2}{E^{A2}(N_2|H_0)}$, which completes the proof for $K = 2$.

Next, we prove the theorem by induction on the number of components K. Assume that the theorem is true for $K-1$ components (where one and only one component is abnormal). Assume that

$$\frac{\pi_1(t_1)w_1}{E^{A2}(N_1|H_0)} \geq \frac{\pi_2(t_1)w_2}{E^{A2}(N_2|H_0)} \geq ... \geq \frac{\pi_K(t_1)w_K}{E^{A2}(N_K|H_0)} \,. \tag{24}$$

Consider an optimal selection rule $\boldsymbol{\phi}^{(j)}$ that selects component j first.

Let

$$\gamma_j(t) = \frac{1}{\pi_j(t)\frac{f_j^{(1)}(\mathbf{y}_j(N_j))}{f_j^{(0)}(\mathbf{y}_j(N_j))} + 1 - \pi_j(t)} \,. \tag{25}$$

Note that when the IDS completes to test component j, the other components update their beliefs according to:

$$\pi_k(t_2) = \gamma_j(t_1)\pi_k(t_1) \;,\; \forall k \neq j \;. \tag{26}$$

The expected weighted sum of completion times achieved by $\boldsymbol{\phi}^{(j)}$ given the outcome (at time t_2) by testing component j (i.e., given the observations vector $\mathbf{y}_j(N_j)$) is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j)\right\} \\ &= \pi_j(t_2)w_j N_j \\ &+ (1-\pi_j(t_2)) \times \\ &\quad \mathbf{E}\left\{\sum_{k=1,k\neq j}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j), j\in\mathcal{H}_0\right\} \,.\end{aligned} \tag{27}$$

Let

$$\widetilde{C}_k = C_k - N_j \quad \forall k \neq j \tag{28}$$

be the modified completion time, defined as the completion time from $t = N_j + 1$ until testing component k is completed. Thus, we can rewrite (27) as:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j)\right\} \\ &= \sum_{k=1}^{K} \pi_k(t_2) w_k N_j \\ &+ (1-\pi_j(t_2)) \times \\ &\quad \mathbf{E}\left\{\sum_{k=1,k\neq j}^{K} w_k \widetilde{C}_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j), j\in\mathcal{H}_0\right\} \,.\end{aligned} \tag{29}$$

The term $\sum_{k=1}^{K} \pi_k(t_2) w_k N_j$ in (29) follows since,

$$\begin{aligned}&\Pr\left(k\in\mathcal{H}_1 \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j), j\in\mathcal{H}_0\right) \\ &= \frac{\Pr\left(k\in\mathcal{H}_1, j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j),\right)}{\Pr\left(j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j),\right)} \\ &= \frac{\Pr\left(k\in\mathcal{H}_1 \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j),\right)}{\Pr\left(j\in\mathcal{H}_0 \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j),\right)} = \frac{\pi_k(t_2)}{1-\pi_j(t_2)} \triangleq \widetilde{\pi}_k(t_2) \,.\end{aligned} \tag{30}$$

Minimizing

$$\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j)\right\} \tag{31}$$

at time t_2, requires to minimize

$$\mathbf{E}\left\{\sum_{k=1,k\neq j}^{K} w_k \widetilde{C}_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid \boldsymbol{\phi}^{(j)}, \mathbf{y}_j(N_j), j\in\mathcal{H}_0\right\} \quad (32)$$

in (29).

Note that (32) is the expected weighted sum of completion times for $K-1$ components (where one and only one component is abnormal) starting at time $t=t_2=N_j+1$, with prior probability $\widetilde{\pi}_k(t_2)=\frac{\pi_k(t_2)}{1-\pi_j(t_2)}$ for component $k\neq j$ to be abnormal. By the induction hypothesis, for any optimal selection rule $\boldsymbol{\phi}^{(j)}$ that selects component j first, arranging the last $K-1$ components with decreasing order of $\widetilde{\pi}_k(t_2)w_k/E^{A2}(N_k|H_0)$ (and testing them by the SPRT) minimizes (32).

Since

$$\widetilde{\pi}_k(t_2)=\frac{\gamma_j(t_1)}{1-\pi_j(t_2)}\pi_k(t_1)\ \ \forall k\neq j, \quad (33)$$

then

$$\frac{\widetilde{\pi}_1(t_2)w_1}{E^{A2}(N_1|H_0)}\geq\frac{\widetilde{\pi}_2(t_2)w_2}{E^{A2}(N_2|H_0)}\geq\cdots\geq\frac{\widetilde{\pi}_{j-1}(t_2)w_{j-1}}{E^{A2}(N_{j-1}|H_0)} \geq\frac{\widetilde{\pi}_{j+1}(t_2)w_{j+1}}{E^{A2}(N_{j+1}|H_0)}\geq\cdots\geq\frac{\widetilde{\pi}_K(t_2)w_K}{E^{A2}(N_K|H_0)}. \quad (34)$$

Thus, the last $K-1$ components have to be selected in decreasing order of $\pi_k(t_1)w_k/E^{A2}(N_k|H_0)$ and tested by the SPRT.

Hence, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})\right\}\\ &=\pi_j(t_1)w_j\left(E'(N_j|H_1,t_1)\right)\\ &+\sum_{k=1,k\neq j}^{K}\left[\pi_k(t_1)w_k\times\right.\\ &\left.\left(E'\left(N_j|H_0,t_1\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A2}\left(N_i|H_0\right)\right)+E^{A2}\left(N_k|H_1\right)\right)\right].\end{aligned} \quad (35)$$

First, note that the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}'(t),\boldsymbol{\delta}'(t),\boldsymbol{\phi}^{(j)})$ can be further reduced for all j by minimizing the expected sample size $E'(N_j|H_i,t_1)$ for $i=0,1$, which achieved by $(\tau_j^{A2},\delta_j^{A2})$. Therefore, an optimal solution has to be $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j)})$ for an optimal selection rule $\boldsymbol{\phi}^{(j)}$. Thus, in the following we consider solutions of the form $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi})$.

Next, by contradiction, consider an optimal selection rule $\boldsymbol{\phi}^{(j\neq1)}$ that selects component $j\neq1$ first. Therefore, $\boldsymbol{\phi}^{(j\neq1)}$ selects the components by the following order:

$$j,1,2,...,j-1,j+1,...,K.$$

As a result, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})\right\}\\ &=\pi_j(t_1)w_j\left(E^{A2}(N_j|H_1)\right)\\ &+\pi_1(t_1)w_1\left[E^{A2}\left(N_j|H_0\right)+E^{A2}\left(N_1|H_1\right)\right]\\ &+\sum_{k=2,k\neq j}^{K}\left[\pi_k(t_1)w_k\times\right.\\ &\left.\left(E^{A2}\left(N_j|H_0\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A2}\left(N_i|H_0\right)\right)+E^{A2}\left(N_k|H_1\right)\right)\right].\end{aligned} \quad (36)$$

We use the interchange argument to prove the theorem. Consider a selection rule $\boldsymbol{\phi}^{(1)}$ that selects component 1 first followed by components $j,2,3,j-1,j+1,...,K$. Similar to (36), the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})\right\}\\ &=\pi_1(t_1)w_1\left(E^{A2}(N_1|H_1)\right)\\ &+\pi_j(t_1)w_j\left[E^{A2}\left(N_1|H_0\right)+E^{A2}\left(N_j|H_1\right)\right]\\ &+\sum_{k=2,k\neq j}^{K}\left[\pi_k(t_1)w_k\times\right.\\ &\left.\left(E^{A2}\left(N_j|H_0\right)+\left(\sum_{i=1,i\neq j}^{k-1}E^{A2}\left(N_i|H_0\right)\right)+E^{A2}\left(N_k|H_1\right)\right)\right].\end{aligned} \quad (37)$$

By comparing (36) and (37), it can be verified that:

$$\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(1)})\right\} \leq \mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A2},\boldsymbol{\delta}^{A2},\boldsymbol{\phi}^{(j\neq1)})\right\}$$

since that

$$\frac{\pi_1(t_1)w_1}{E^{A2}(N_1|H_0)}\geq\frac{\pi_j(t_1)w_j}{E^{A2}(N_j|H_0)}.$$

The expected weighted sum of completion times can be reduced by selecting component 1 first followed by component j, which contradicts the optimality of $\boldsymbol{\phi}^{(j\neq1)}$. Hence, at time t_1 selecting component 1 minimizes the expected weighted sum of completion times. We have already proved that selecting the last $K-1$ components in decreasing order of $\pi_k(t_1)w_k/E^{A2}(N_k|H_0)$ minimizes the objective function, which completes the proof. ■

VI. Localization of Anomaly Under Uncertainty

In the previous sections, we focused on the simple hypotheses case, where the distribution under both hypotheses

are completely known. For this case, the SPRT was applied in Algorithms $1, 2$ to solve (3). However, in numerous cases under the adversary model, there is uncertainty in the observation distribution (in particular when the component is in an abnormal state). Therefore, in this section we extend our results to the case of composite hypotheses, where there is uncertainty in the distribution parameters.

Let $\boldsymbol{\theta}_k$ be a vector of unknown parameters of component k. The observations $\{y_k(i)\}_{i\geq 1}$ are drawn from a common distribution $f\left(y|\boldsymbol{\theta}_k\right)$, $\boldsymbol{\theta}_k \in \Theta_k$, where Θ_k is the parameters space of component k. If component k is in a healthy state, then $\boldsymbol{\theta}_k \in \Theta_k^{(0)}$; if component k is abnormal, then $\boldsymbol{\theta}_k \in (\Theta \backslash \Theta_k^{(0)})$.

Let $\Theta_k^{(0)}$, $\Theta_k^{(1)}$ be disjoint subsets of Θ_k, where $I_k = \Theta \backslash (\Theta_k^{(0)} \cup \Theta_k^{(1)}) \neq \emptyset$ is an indifference region[3]. When $\boldsymbol{\theta}_k \in I_k$, the detector is indifferent regarding the state of component k. Hence, there are no constraints on the error probabilities for all $\boldsymbol{\theta}_k \in I_k$. The hypothesis testing regarding component k is to test

$$\boldsymbol{\theta}_k \in \Theta_k^{(0)} \quad \text{against} \quad \boldsymbol{\theta}_k \in \Theta_k^{(1)}.$$

Narrowing I_k has the price of increasing the sample size.
Let

$$\begin{aligned} \hat{\boldsymbol{\theta}}_k(n) &= \arg \max_{\theta_k \in \Theta_k} f\left(\mathbf{y}_k(n)|\boldsymbol{\theta}_k\right), \\ \hat{\boldsymbol{\theta}}_k^{(i)}(n) &= \arg \max_{\theta_k \in \Theta_k^{(i)}} f\left(\mathbf{y}_k(n)|\boldsymbol{\theta}_k\right), \end{aligned} \tag{38}$$

be the Maximum-Likelihood Estimates (MLEs) of the parameters over the parameters space Θ_k, $\Theta_k^{(i)}$ at stage n, respectively.

In contrast to the SPRT (for the simple hypotheses case), the theory of sequential tests of composite hypotheses does not provide optimal performance in terms of minimizing the expected sample size under given error constraints. Nevertheless, asymptotically optimal performance can be obtained as the error probability approaches zero.

First, we provide an overview of existing sequential tests for composite hypotheses which are relevant to our problem. Next, we apply these techniques to solve (2).

A. Existing Sequential Tests for Composite Hypothesis Testing

The key idea of sequential tests of composite hypotheses, discussed in this section, is to use the estimated parameters to perform a one-sided sequential test to reject H_0 and a one-sided sequential test to reject H_1. Note that these techniques were introduced for a single process. However, in this paper we apply sequential tests for K components. Thus, we use the subscript k to denote the component index.

1) Sequential Generalized Likelihood Ratio Test (SGLRT): We refer to sequential tests that use the Generalized Likelihood Ratio (GLR) statistics [32] as the SGLRT.
For $i = 0, 1$, let

$$L_k^{(i),GLR}(n) = \log \frac{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k(n))}{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k^{(i)}(n))} \tag{39}$$

be the GLR statistics used to reject hypothesis H_i at stage n. Let

$$N_k^{(i)} = \inf\left\{ n \ : L_k^{(i),GLR}(n) \geq B_k^{(i)} \right\}, \tag{40}$$

be the stopping rule used to reject hypothesis H_i. $B_k^{(i)}$ is the boundary value.
For each component k, the IDS stops sampling when $N_k = \min\left\{N_k^{(0)}, N_k^{(1)}\right\}$. If $N_k = N_k^{(0)}$, component k is declared as abnormal (i.e., H_0 is rejected). If $N_k = N_k^{(1)}$, component k is declared as normal (i.e., H_0 is accepted).

The SGLRT was first studied by Schwartz [24] for a one-parametric exponential family, who assigned a cost of c for each observation and a loss function for wrong decision. It was shown that setting $B_k^{(i)} = \log(c_k^{-1})$ asymptotically minimizes the Bayes risk as c_k approaches zero. Further refinement was studied by Lai [27], [29], who set a time-varying boundary value $B_k^{(i)} \sim \log((nc_k)^{-1})$. Lai showed that for a multivariate exponential family this scheme asymptotically minimizes both the Bayes risk and the expected sample size subject to error constraints as c_k approaches zero [29].

2) Sequential Adaptive Likelihood Ratio Test (SALRT): We refer to sequential tests that use the Adaptive Likelihood Ratio (ALR) statistics as the SALRT.
For $i = 0, 1$, let

$$L_k^{(i),ALR}(n) = \log \frac{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k(r-1))}{\prod_{r=1}^{n} f(y_k(r)|\hat{\boldsymbol{\theta}}_k^{(i)}(n))} \tag{41}$$

be the ALR statistics used to reject hypothesis H_i at stage n. Let

$$N_k^{(i)} = \inf\left\{ n \ : L_k^{(i),ALR}(n) \geq B_k^{(i)} \right\}, \tag{42}$$

be the stopping rule used to reject hypothesis H_i.
For each component k, the IDS stops sampling when $N_k = \min\left\{N_k^{(0)}, N_k^{(1)}\right\}$. If $N_k = N_k^{(0)}$, component k is declared as abnormal (i.e., H_0 is rejected). If $N_k = N_k^{(1)}$, component k is declared as normal (i.e., H_0 is accepted).

The SALRT was first introduced by Robbins and Siegmund [25], [26] to design power-one sequential tests. Pavlov used it to design asymptotically (as the error probability approaches zero) optimal (in terms of minimizing the expected sample size subject to error constraints) tests for composite hypothesis testing of multivariate exponential family [28]. Tartakovsky shows asymptotically optimal performance for a more general multivariate family of distributions [30].

The advantage of using the SALRT is that setting $B_k^{(0)} = \log \frac{1}{\alpha_k}$, $B_k^{(1)} = \log \frac{1}{\beta_k}$ satisfies the error probability constraints in (3). However, such simple setting can not be applied to the SGLRT. Thus, implementing the SALRT is much simpler than implementing the SGLRT. The disadvantage of using the SALRT is that poor early estimates (for small number of observations) can never be revised even though one has a large number of observations. Thus, generally, the SGLRT outperforms the SALRT in terms of minimizing the expected sample size for given type I and type II errors.

[3]The assumption of an indifference region is widely used in the theory of sequential testing of composite hypotheses to derive asymptotically optimal performance. Nevertheless, in some cases this assumption can be removed. For more details, the reader is referred to [27].

TABLE III

ALGORITHM 3 FOR THE INDEPENDENT MODEL UNDER UNCERTAINTY

1. arrange the components in decreasing order of $\pi_k(t_1)w_k/E(N_k)$
2. for $k = 1, ..., K$ components do:
3. perform SALRT/SGLRT for component k, with $P_k^{FA} \leq \alpha_k$, $P_k^{MD} \leq \beta_k$
4. end for

TABLE IV

ALGORITHM 4 FOR THE EXCLUSIVE MODEL UNDER UNCERTAINTY

1. arrange the components in decreasing order of $\pi_k(t_1)w_k/E(N_k|H_0)$
2. for $k = 1, ..., K$ components do:
3. perform SALRT/SGLRT for component k, with $P_k^{FA} \leq \alpha_k$, $P_k^{MD} \leq \beta_k$
4. end for

B. The Proposed Solutions for the Independent and Exclusive Models

In this section we modify Algorithms $1, 2$, given in Tables I, II to take into account the uncertainty in the model of the adversary. Based on the solution to the two-stage optimization problem, we propose Algorithm 3 and 4 to solve (2) for the independent and exclusive models under uncertainty, respectively. The algorithms are presented in Tables III, IV. The required modification is in step 3 of both algorithms. Under uncertainty, one should perform SGLRT or SALRT, as discussed in the previous section, instead of the SPRT.

Remark 3*:* Implementing Algorithms $3, 4$ requires to compute the expected sample size $E(N_k|H_i)$ for all $k = 1, 2, ..., K$ for $i = 0, 1$, achieved by the SGLRT or the SALRT. In general, it is difficult to obtain a closed-form expressions to the exact value of $E(N_k|H_i)$. However, we can use the asymptotic property of the tests to obtain a closed-form approximation to $E(N_k|H_i)$, which approaches the exact expected sample size as the error probability approaches zero.
For every $i = 0, 1$, let

$$D_k(\boldsymbol{\theta}_k||\boldsymbol{\lambda}) = E_{\boldsymbol{\theta}_k}\left(\log\frac{f(y_k(1)|\boldsymbol{\theta}_k)}{f(y_k(1)|\boldsymbol{\lambda})}\right) \tag{43}$$

be the KL divergence between the real value of $\boldsymbol{\theta}_k$ and $\boldsymbol{\lambda}$, where the expectation is taken with respect to $f(y|\boldsymbol{\theta}_k)$, and let

$$D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(i)}) = \inf_{\boldsymbol{\lambda}\in\Theta_k^{(i)}} D_k(\boldsymbol{\theta}_k||\boldsymbol{\lambda}) \ . \tag{44}$$

Let $P^{(i)}(\boldsymbol{\theta}_k)$ be a prior distribution on $\boldsymbol{\theta}_k$ under hypothesis H_i at component k. Then, as $P_k^{FA} \to 0, P_k^{MD} \to 0$, the expected sample size is given by:

$$\begin{aligned} E(N_k|H_0) &\sim \int_{\boldsymbol{\theta}_k\in\Theta_k^{(0)}} \frac{\log B_k^{(1)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(1)})} dP^{(0)}(\boldsymbol{\theta}_k) \ , \\ E(N_k|H_1) &\sim \int_{\boldsymbol{\theta}_k\in\Theta_k^{(1)}\cup I_k^{(1)}} \frac{\log B_k^{(0)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(0)})} dP^{(1)}(\boldsymbol{\theta}_k) \\ &\quad + \int_{\boldsymbol{\theta}_k\in I_k^{(0)}} \frac{\log B_k^{(1)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(1)})} dP^{(1)}(\boldsymbol{\theta}_k) \ , \end{aligned} \tag{45}$$

where $I_k^{(0)}, I_k^{(1)}$ are disjoint subsets of I_k and $I_k = I_k^{(0)} \cup I_k^{(1)}$. For all $\boldsymbol{\theta}_k \in I_k^{(i)}$ we have $\frac{\log B_k^{(j)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(j)})} \leq \frac{\log B_k^{(i)}}{D_k^*(\boldsymbol{\theta}_k||\Theta_k^{(i)})}$ for $i, j = 0, 1$.
At each time t, the expected sample size required to make a decision regarding the state of component k is given by:

$$E(N_k) = \pi_k(t)E(N_k|H_1) + (1 - \pi_k(t))E(N_k|H_0) \ , \tag{46}$$

which can be well approximated for small error probability using (45). *Remark* 4*:* In numerous cases, uncertainty is associated with abnormal state solely, where the distribution under normal state is completely known. In these cases, evaluating $E(N_k)$ to implement Algorithm 3 depends on the prior distribution on $\boldsymbol{\theta}_k \in \Theta\backslash\Theta_k^{(0)}$, while evaluating $E(N_k|H_0)$ to implement Algorithm 4 does not.

C. Asymptotic Optimality of Algorithms $3, 4$

In what follows we show that Algorithms $3, 4$ are asymptotically optimal in terms of minimizing the objective function subject to the error constraints (2) as the error probability approaches zero. When deriving asymptotic we assume that $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k such that the asymptotic optimality property in terms of minimizing the expected sample size subject to the error constraints holds for each single process for both SGLRT and SALRT, as discussed in Section VI-A[4].

Theorem 5: Consider the independent model under uncertainty. Let $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ *be the optimal solution to (2). Let* $(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})$ *be the solution achieved by Algorithm* 3*. Then, as* $P_k^{FA} \to 0, P_k^{MD} \to 0$ *for all* k*, we obtain:*

$$\begin{aligned} &\boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k|(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})\right\} \\ &\qquad \sim \boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k|(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)\right\} \end{aligned} \tag{47}$$

Proof: For every k, let $E^*(N_k|H_i)$ be the minimal expected

[4]Asymptotic optimality for a single process is guaranteed for an exponential family of distributions when $\log P_k^{FA} \sim \log P_k^{MD} \sim \log B^{-1}$ (which is satisfied by setting $B_k^{(i)} = d_k^{(i)} B$ for $i = 0, 1$ for some positive constants $d_k^{(i)}$ and letting B approach infinity) under some weak conditions on the parameter distribution. Nevertheless, more general results can be obtained in some cases. For more details, the reader is referred to Section VI-A and references therein.

sample size that can be achieved by any sequential test, such that error constraints are satisfied. Let $E^{A3}(N_k|H_i)$ be the expected sample size achieved by Algorithm 3, such that error constraints are satisfied. Clearly, $E^*(N_k|H_i) \leq E^{A3}(N_k|H_i)$ for all k, for $i = 0, 1$.
Assume that

$$\frac{\pi_1(t_1)w_1}{E^*(N_1)} \geq \frac{\pi_2(t_1)w_2}{E^*(N_2)} \geq ... \geq \frac{\pi_K(t_1)w_K}{E^*(N_K)}. \tag{48}$$

Similar to the proof of Theorem 2, it can be verified that the optimal solution to (2) is given by selecting the components by the following order: $1, 2, ..., K$, where the components are tested by a sequential test that achieves expected sample size $E^*(N_k|H_i)$ for all k, for $i = 0, 1$. Therefore, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)\right\}\\ &= \sum_{k=1}^{K} \pi_k(t_1) w_k \left[\left(\sum_{i=1}^{k-1} E^*\left(N_i\right)\right) + E^*\left(N_k|H_1\right)\right].\end{aligned} \tag{49}$$

By the asymptotic optimality property of the SALRT/SGLRT for a single process (used in Algorithm 3), it follows that $E^{A3}(N_k|H_i) \sim E^*(N_k|H_i)$ for all k, for $i = 0, 1$ as $P_k^{FA} \to 0, P_k^{MD} \to 0$. As a result, for sufficiently small error probabilities, the solution $(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})$ is given by selecting the components by the following order: $1, 2, ..., K$, where the components are tested by an asymptotically optimal sequential test that achieves expected sample size $E^{A3}(N_k|H_i)$ for all k, for $i = 0, 1$. Therefore, the expected weighted sum of completion times achieved by $(\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})$ is given by:

$$\begin{aligned}&\mathbf{E}\left\{\sum_{k=1}^{K} w_k C_k \mathbf{1}_{\{k\in\mathcal{H}_1\}} \mid (\boldsymbol{\tau}^{A3}, \boldsymbol{\delta}^{A3}, \boldsymbol{\phi}^{A3})\right\}\\ &= \sum_{k=1}^{K} \pi_k(t_1) w_k \left[\left(\sum_{i=1}^{k-1} E^{A3}\left(N_i\right)\right) + E^{A3}\left(N_k|H_1\right)\right].\end{aligned} \tag{50}$$

Since $E^{A3}(N_k|H_i) \sim E^*(N_k|H_i)$ for $i = 0, 1$ as $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k, the theorem follows. ■

Theorem 6: Consider the exclusive model under uncertainty. Let $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ be the optimal solution to (2). Let $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ be the solution achieved by Algorithm 4. Then, as $P_k^{FA} \to 0, P_k^{MD} \to 0$ for all k, we obtain:

$$\begin{aligned}&\boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k | (\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})\right\}\\ &\qquad \sim \boldsymbol{E}\left\{\sum_{k\in\mathcal{H}_1} w_k C_k | (\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)\right\}\end{aligned} \tag{51}$$

Proof: The structure of the proof is similar to the proof of Theorem 5. Hence, we provide a sketch of the proof, using similar notations used in the proof of Theorem 5. Similar to the proof of Theorem 4, it can be verified that the optimal solution to (2) is given by selecting the components in decreasing order of $\pi_k(t_1)w_k/E^*(N_k|H_0)$, where the components are tested by a sequential test that achieves expected sample size $E^*(N_k|H_i)$ for all k, for $i = 0, 1$. By the asymptotic optimality property for a single process of the SALRT/SGLRT (used in Algorithm 4), it follows that $E^{A4}(N_k|H_i) \sim E^*(N_k|H_i)$ for all k, for $i = 0, 1$ as $P_k^{FA} \to 0, P_k^{MD} \to 0$. As a result, for sufficiently small error probabilities, the solution $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ is given by selecting the components in decreasing order of $\pi_k(t_1)w_k/E^*(N_k|H_0)$, where the components are tested by an asymptotically optimal sequential test that achieves expected sample size $E^{A4}(N_k|H_i)$ for all k, for $i = 0, 1$. Similar to the proof of Theorem 5, comparing the objective functions achieved by $(\boldsymbol{\tau}^*, \boldsymbol{\delta}^*, \boldsymbol{\phi}^*)$ and $(\boldsymbol{\tau}^{A4}, \boldsymbol{\delta}^{A4}, \boldsymbol{\phi}^{A4})$ proves the theorem. ■

VII. Applications and Numerical Examples

In this section, we provide applications and numerical examples to illustrate the performance of the algorithms. Assume that an intruder tries to launch a Denial of Service (DoS) or Reduction of Quality (RoQ) attacks by sending a large number of packets to a component (which can be a relay node in this application). DoS attacks rely on overwhelming the component with useless traffic that constantly exceeds its capacity so to make it unavailable for its intended use. On the other hand, RoQ attacks inflict damage on the component, while keeping a low profile to avoid detection. RoQ attacks do not cause denial of service.

To detect such attacks, the IDS performs a traffic-based anomaly detection. It monitors the traffic at each component to decide whether a component is compromised. Roughly speaking, if the actual arrival rate is significantly higher than the arrival rate under normal state, then the IDS should declare that the component is in an abnormal state. Similar traffic-based detection techniques were proposed in [7], [12] for different models, considering a single process without switching to other nodes.

For each component k, we assume that packets arrive according to a Poisson process with rate $\theta^{(k)}$, which is generally considered to be a good model in a queuing theory analysis [33]. When component k is tested, the IDS collects an observation $y_k(n) \in \mathbb{N}_0$ every time unit, which represents the number of packets that arrived in the interval $(n-1, n)$. Assume that the IDS considers component k as normal if $\theta_k \leq \theta_k^{(0)}$, and tests $\theta_k \leq \theta_k^{(0)}$ against $\theta_k \geq \theta_k^{(1)}$ (i.e., $I_k = \{\theta_k | \theta_k^{(0)} < \theta_k < \theta_k^{(1)}\}$ is the indifference region).

We set $w_k = \theta_k^{(0)}$. Under this setting, the objective function represents the total expected number of failed packets in the network during DoS attacks. Thus, the optimization problem can be observed as minimizing the maximal damage to the network in terms of packet-loss. Furthermore, this setting prioritizes components with higher normal traffic to reduce the delay caused by RoQ attacks.

A. Detection Under Simple Hypotheses

In this section, we consider the case where the parameters $\theta_k = \theta_k^{(0)}$ under normal state and $\theta_k = \theta_k^{(1)}$ under abnormal state are known to the IDS. To implement Algorithms $1, 2$

(which are optimal in this scenario for the independent and exclusive model, respectively), we need to compute the LR (or the log-LR) between the hypotheses, defined in (4), and the expected sample sizes under the hypotheses, which can be well approximated by (9).
Let

$$\Lambda_k(n) = \log L_k(n) \tag{52}$$

be the Log-Likelihood Ratio (LLR) between the two hypotheses of component k at stage n, where $L_k(n)$ is defined in (4). After algebraic manipulations, it can verified that the LLR is given by:

$$\Lambda_k(n) = -n\left(\theta_k^{(1)} - \theta_k^{(0)}\right) + \log\left(\theta_k^{(1)}/\theta_k^{(0)}\right)\sum_{i=1}^{n} y_k(i)\ . \tag{53}$$

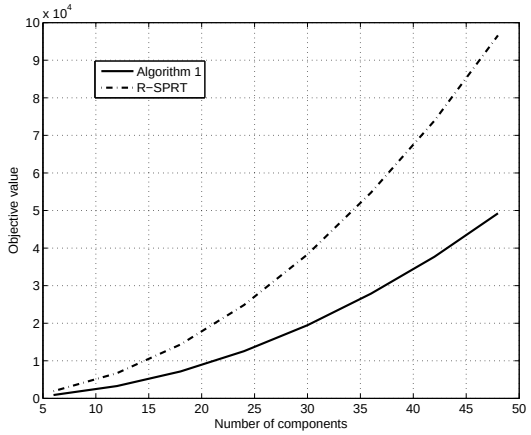
It can be verified that the KL divergence between the hypotheses H_i and H_j, defined in (8), is given by:

$$D_k(i||j) = \theta_k^{(j)} - \theta_k^{(i)} + \theta_k^{(i)}\log\left(\theta_k^{(i)}/\theta_k^{(j)}\right)\ . \tag{54}$$

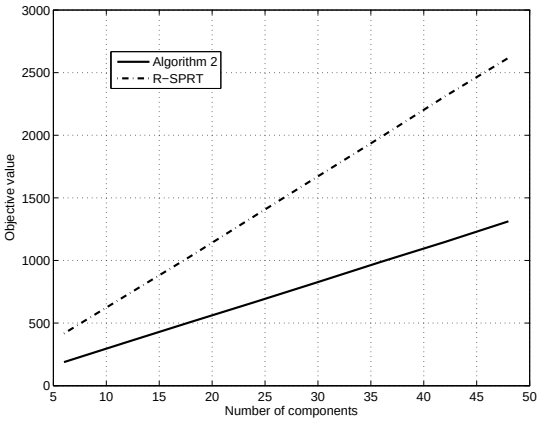
Substituting (54) in (9) obtains the required approximation to the expected sample size.

Next, we provide numerical example to illustrate the performance of the algorithms. We compared three schemes: a Random selection SPRT (R-SPRT), where a series of SPRTs are performed until all the components are tested in a random order, and the proposed Algorithms $1, 2$, which are optimal for the independent and exclusive models, respectively.

Let $\delta_K = (100-10)/(K-1)$. We set $w_k = \theta_k^{(0)} = 10 + (k-1)\delta_K$ and $\theta_k^{(1)} = 1.5\cdot\theta_k^{(0)}$. The error constraints were set to $P_k^{FA} = P_k^{MD} = 10^{-2}$ for all k. For the independent and exclusive models, we set $\pi_k = 0.8$ and $\pi_k = 1/K$ for all k, respectively. The performance of Algorithm 1 and Algorithm 2 are presented in Fig. 1(a) and 1(b) under the independent and exclusive models, respectively, as compared to the R-SPRT. It can be seen that the proposed Algorithms save roughly 50% of the objective value as compared to the R-SPRT under both the independent and exclusive model scenarios.



(a) An independent model scenario.



(b) An exclusive model scenario.

Fig. 1. Objective value as a function of the number of components under the independent and exclusive models.

B. Detection Under Uncertainty

In this section, we consider the case of composite hypotheses, where there is uncertainty in the distribution parameters (in particular when the component is in an abnormal state), as discussed in Section VI. To implement Algorithms $3, 4$ (which are asymptotically optimal in this scenario for the independent and exclusive model, respectively), we need to compute the GLR or ALR statistics between the hypotheses, defined in (39), (41) and the expected sample sizes under the hypotheses, which can be well approximated by (45). The MLEs of the parameters over the parameter space Θ_k, $\Theta_k^{(i)}$ are given by the sample mean and the boundary of the alternative parameter space, respectively. As a result, substituting:

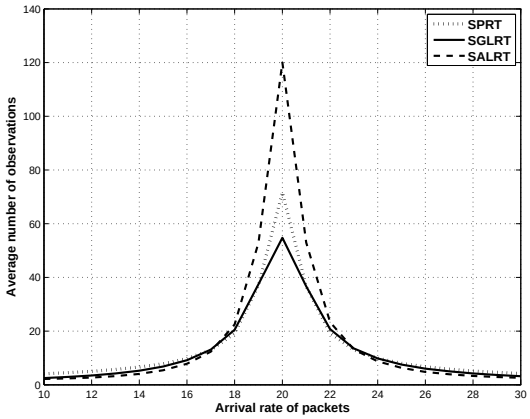
$$\begin{aligned}\hat{\theta}_k(n) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_k(i),\\ \hat{\theta}_k^{(i)}(n) &= \theta_k^{(i)},\end{aligned} \tag{55}$$

in (39), (41) yields the GLR and ALR statistics, respectively. The KL divergence between the real value of θ_k and the parameter space $\Theta_k^{(i)}$ is given by:

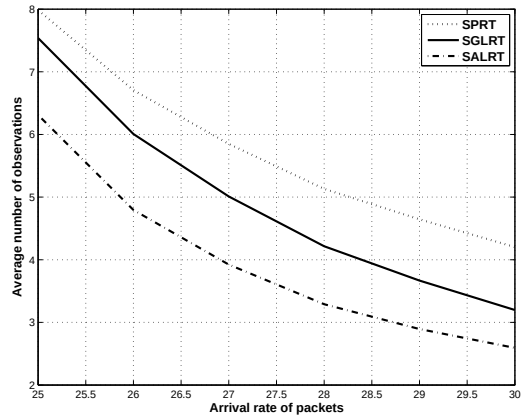
$$D_k^*(\theta_k||\Theta_k^{(i)}) = \theta_k^{(i)} - \theta_k + \theta_k\log\left(\theta_k/\theta_k^{(i)}\right)\ . \tag{56}$$

Substituting (56) in (45) yields the approximate expected sample size.

Next, we provide numerical example to illustrate the performance of the algorithms under uncertainty. We simulated a network with homogenous components (i.e., any selection rule is optimal). We compared three schemes: R-SPRT, and Algorithms 3 or 4 (which achieve the same performance in this case) using the SALRT and the SGLRT, discussed in section VI-A. We set $\theta_k^{(0)} = 19$, $\theta_k^{(1)} = 21$. Under uncertainty, the IDS considers component k as normal if $\theta_k \le \theta_k^{(0)}$, and tests $\theta_k \le \theta_k^{(0)}$ against $\theta_k \ge \theta_k^{(1)}$ (i.e., $I_k = \{\theta_k | 19 < \theta_k < 21\}$ is the indifference region). To implement the SGLRT, we set the cost per observation $c = 10^{-3}$. According to the assigned cost, we obtained the following error probability constraints for all k: $P_k^{FA} \le 0.026$ for all $\theta^{(k)} \le 19$ and $P_k^{MD} \le 0.03$ for all $\theta^{(k)} \ge 21$. We do not restrict the detector's performance for $19 < \theta^{(k)} < 21$ (Note that narrowing the indifference region has the price of increasing the sample size). In Fig. 2 we show the average number of observations required for detection as a function of $\theta^{(k)}$. As expected, for $\theta_k = 19$ and $\theta_k = 21$ the R-



(a) Average number of observations as a function of θ



(b) Average number of observations as a function of θ

Fig. 2. Average number of observations as a function of the arrival rate of packets (denoted by θ).

SPRT requires lower sample size as compared to the proposed schemes. On the other hand, it can be seen that for most values of θ the SGLRT and the SALRT require lower sample size as compared to the R-SPRT. The SALRT performs the worst for $18 < \theta_k < 22$, and performs the best for $\theta_k \notin (18, 22)$, roughly. The SGLRT obtains the best average performance. It can be seen that for large values of θ_k the anomaly is detected very quickly, since the distance between the hypotheses increases. This result confirms that DoS attacks are much easier to detect as compared to RoQ attacks.

VIII. Conclusion

The problem of quickest localization of anomaly in a resource-constrained cyber network was investigated. Due to resource constraints, only one component can be probed at each time. The observations are random realizations drawn from two different distributions depending on whether the component is normal or anomalous. The problem was formulated as a priority-based constrained optimization problem. Components with higher priorities in an abnormal state should be fixed before components with lower priorities to reduce the overall damage to the network. The objective is to minimize the expected weighted sum of completion times subject to error probability constraints. We considered two different anomaly models: the independent model in which each component can be abnormal independent of other components, and the exclusive model in which there is one and only one abnormal component. For the simple hypotheses case, we derived optimal algorithms for both independent and exclusive models. For the composite hypotheses case, we derived asymptotically (as the error probability approaches zero) optimal algorithms for both independent and exclusive models. These optimal algorithms have low-complexity.

The algorithms developed throughout this paper can be applied to other models of anomaly detection as well. We can modify the proposed algorithms to any detection scheme that performs a series of tests until all the components are tested. The required modification is in step 3 of the algorithms, where the SPRT/SALRT/SGLRT are replaced by any given test. As a result, the modified algorithms minimize the objective function among all the algorithms that perform the given test.

References

[1] T. F. Lunt, "A survey of intrusion detection techniques," *Computers & Security*, vol. 12, no. 4, pp. 405–418, 1993.

[2] A. Murali and M. Rao, "A survey on intrusion detection approaches," in *IEEE International Conference on Information and Communication Technologies (ICICT)*, pp. 233–240, 2005.

[3] K. C. Gross and W. Lu, "Early detection of signal and process anomalies in enterprise computing systems," in *Proc. of IEEE International Conference on Machine Learning and Applications*, 2002.

[4] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[5] J. Jung, V. Paxson, A. W. Berger, and H. Balakrishnan, "Fast portscan detection using sequential hypothesis testing," in *Proceedings IEEE Symposium on Security and Privacy*, pp. 211–225, 2004.

[6] J. Jung, S. E. Schechter, and A. W. Berger, "Fast detection of scanning worm infections," in *Proceeding Inter. Symposium on Recent Advances in Intrusion Detection (RAID)*, pp. 59–81, 2004.

[7] I. Onat and A. Miri, "An intrusion detection system for wireless sensor networks," in *IEEE International Conference on Wireless And Mobile Computing, Networking And Communications*, vol. 3, pp. 253–259, 2005.

[8] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blazek, and H. Kim, "A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3372–3382, 2006.

[9] G. Androulidakis, V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, and V. Maglaris, "Understanding and evaluating the impact of sampling on anomaly detection techniques," in *IEEE Military Communications Conference (MILCOM)*, pp. 1–7, 2006.

[10] T. Van Phuong, L. Hung, S. Cho, Y. K. Lee, and S. Lee, "An anomaly detection algorithm for detecting attacks in wireless sensor networks," *Intelligence and Security Informatics*, pp. 735–736, 2006.

[11] T. He and L. Tong, "Detecting encrypted stepping-stone connections," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1612–1623, 2007.

[12] V. B. Misic and J. Begum, "Evaluating the feasibility of traffic-based intrusion detection in an 802.15.4 sensor cluster," in *IEEE International Conference on Advanced Information Networking and Applications*, pp. 619–624, 2007.

[13] A. A. Cárdenas, S. Radosavac, and J. S. Baras, "Evaluation of detection algorithms for mac layer misbehavior: theory and experiments," *IEEE/ACM Transactions on Networking*, vol. 17, no. 2, pp. 605–617, 2009.

[14] G. Thatte, U. Mitra, and J. Heidemann, "Parametric methods for anomaly detection in aggregate traffic," *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 512–525, 2011.

[15] K. Liu and Q. Zhao, "Intrusion detection in resource-constrained cyber networks: A restless multi-armed bandit approach," *submitted to IEEE/ACM Transactions on Networking. Avialable at http://arxiv.org/abs/1112.0101*.

[16] K. Cohen, Q. Zhao, and A. Swami, "Quickest localization of anomaly in cyber networks under uncertainty," *Submitted to the Military Communications Conference (MILCOM)*, 2013.

[17] A. Agah, S. K. Das, K. Basu, and M. Asadi, "Intrusion detection in sensor networks: A non-cooperative game approach," in *Proc. IEEE International Symposium on Network Computing and Applications (NCA)*, pp. 343–346, 2004.

[18] R. S. Blum and B. M. Sadler, "Energy efficient signal detection in sensor networks using ordered transmissions," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3229–3235, 2008.

[19] K. Cohen and A. Leshem, "Energy-efficient detection in wireless sensor networks using likelihood ratio and channel state information," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1671–1683, 2011.

[20] Y. R. Tsai and L. C. Lin, "Sequential fusion for distributed detection over BSC channels in an inhomogeneous sensing environment," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 99–102, 2010.

[21] P. Braca, S. Marano, and V. Matta, "Single-transmission distributed detection via order statistics," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 2042–2048, 2012.

[22] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, pp. 117–186, 1945.

[23] A. Wald, "Sequential analysis," *New York: Wiley*, 1947.

[24] G. Schwarz, "Asymptotic shapes of Bayes sequential testing regions," *The Annals of mathematical statistics*, pp. 224–236, 1962.

[25] H. Robbins and D. Siegmund, "A class of stopping rules for testing parametric hypotheses," *Proceeding of the Sixth Berkeley Symposium on Theory of Probability and Math. Statistics*, pp. 37–41, 1972.

[26] H. Robbins and D. Siegmund, "The expected sample size of some tests of power one," *The Annals of Statistics*, pp. 415–436, 1974.

[27] T. L. Lai, "Nearly optimal sequential tests of composite hypotheses," *The Annals of Statistics*, pp. 856–886, 1988.

[28] I. V. Pavlov, "Sequential procedure of testing composite hypotheses with applications to the Kiefer-Weiss problem," *Theory of Probability and Its Applications*, vol. 35, no. 2, pp. 280–292, 1990.

[29] T. L. Lai and L. M. Zhang, "Nearly optimal generalized sequential likelihood ratio tests in multivariate exponential families," *Lecture Notes-Monograph Series*, pp. 331–346, 1994.

[30] A. G. Tartakovsky, "An efficient adaptive sequential procedure for detecting targets," in *IEEE Aerospace Conference Proceedings, 2002*, vol. 4, pp. 1581–1596, 2002.

[31] W. E. Smith, "Various optimizers for single-stage production," *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 59–66, 1956.

[32] S. M. Kay, "Fundamentals of statistical signal processing, Volume II: Detection theory," *Upper Saddle River (New Jersey)*, vol. 7, 1998.

[33] D. P. Bertsekas and R. G. Gallager, *Data networks*. (2nd edition) Prentice Hall, 1992.